

On the Relationship Between Interpretability and Explainability in Machine Learning

Benjamin Leblanc

Université Laval, Québec, Canada
benjamin.leblanc.2@ulaval.ca

Abstract. Interpretability and explainability have gained more and more attention in the field of machine learning as they are crucial when it comes to high-stakes decisions and troubleshooting. Since both provide information about predictors and their decision process, they are often seen as two independent means for one single end. This view has led to a dichotomous literature: explainability techniques designed for complex black-box models, or interpretable approaches ignoring the many explainability tools. In this position paper, we challenge the common idea that interpretability and explainability are substitutes for one another by listing their principal shortcomings and discussing how both of them mitigate the drawbacks of the other. In doing so, we call for a new perspective on interpretability and explainability, and works targeting both topics simultaneously, leveraging each of their respective assets.

Keywords: Interpretability · Explainability · Relationship · Definition.

1 Introduction

Machine learning (ML) is used in a variety of fields with numerous applications, such as image recognition [51], sentiment analysis [79], and language translation [21]. In some areas such as the medical field, ML-assisted predictions or decisions can drastically impact human life. For example, breast cancer [76] can be devastating if not diagnosed in time (or at all). Still, many events made it clear that not understanding the inner workings and decision process of a predictor could lead to unfortunate events: job and loan applications biased toward men [17], mortgage-approval biased toward white applicants [68], higher credit card limits for men [102], etc. As pointed out by Goodman and Flaxman [36]: “If we do not know how ML [predictors] work, we cannot check or regulate them to ensure that they do not encode discrimination against minorities [...], we will not be able to learn from instances in which it is mistaken.”

Two schools of thought seem to have emerged from this need for knowledge: either focusing on explaining black-boxes [81,39], or favoring interpretable predictors while completely discarding explainability [85]. Both approaches contain many flaws: interpretability can’t provide all that there is to know about a predictor [6], all the while being subjective [86], and interpretable predictors are, most of the time, harder to train than predictors in general [27]. On the other

hand, there is an obligation to trust the explanations [10] which, by definition, are necessarily wrong [73,85] and can provide more question than it answers [44]. As it is often understood that explainability reduces the need for interpretability, and *vice versa*, the literature on their practical application and the mitigation of their drawbacks is disjoint.

In this work, we challenge the common belief that interpretability and explainability are substitutes for one another and argue that they actually are complementary, especially in that they mitigate the shortcomings of the other. Our first contribution is met by listing and discussing the principal drawbacks of explainability and interpretability raised in the literature. For our second contribution, we discuss how these limitations either disappear or diminish when considering both explained and interpretable predictors. Throughout the paper, we also show how both concepts differ in their very nature, thus necessarily aiming at different knowledge.

Many works critically discuss interpretability and its relationship with predictive performances [85] or with ML family of predictors [60], and its conceptual differences with explainability [60,86,56,50,19]. Yet, it is the first time, up to our knowledge, that a reconciliation between interpretability and explainability via their complementarity is attempted and pleaded. In doing so, we call for a new perspective on interpretability and explainability and for works targeting both topics simultaneously, leveraging each of their respective assets in practical applications.

2 Defining and Discussing the Concepts at Hand

Cases occur where a lot of definitions are attempted for defining a single concept (such as for interpretability [39,33,23,116,37,32,124,20,30,87,4,66]), many resembling each other. Thus, clearly defining the terms that are used or which paradigm is referred to is necessary for ensuring efficient transfers of ideas and avoiding misunderstandings.

A *task* is an ML problem, whereas a *domain* is a set of tasks sharing similar explanatory variables and response variables. The prediction (decision) process of a predictor corresponds to how it transforms a given input into an output.

The *inputs* of a problem corresponds to the explanatory variables. More broadly, a *feature* could be an input, or could emerge from the interaction between several inputs, but is characterized by being meaningful. Sex, gender, postal code, annual income, and number of cars are features, whereas pixels and sound files are not.

We distinguish ML *models*, *algorithms*, and *predictors*. A predictor simply is a mathematical function. An algorithm tunes the *parameters* and the form of a predictor for it to become better, given criteria, at tackling a given task. Algorithms can also build a predictor from scratch [57,11,64]. In the machine learning literature, the term model is often used loosely to refer interchangeably to an algorithm, a predictor, or a family of predictors. In the following, we retain the last definition: a machine learning model is a family of predictors. For

example, a linear model is the space of predictors outputting a linear function of its inputs.

2.1 Interpretability

We do not aim at attempting a new definition of interpretability in machine learning: we rather refer to what preexisting paradigm we stand in. Inspired by Rudin [86], we consider that a degree of *interpretability* of a predictor on a given task is proportional to the capacity of a given person to understand its decision process simply and only by considering the predictor in itself.

Just as predictive performances, many metrics could be used for measuring such a degree. Here, a predictor whose degree of interpretability is high will be simply called *interpretable*, whereas when it is really low, a *black-box*. A person evaluating the interpretability of a predictor will be called a *judge*. Since the decision process of a predictor (both the inputs and the output) encapsulates the notion of *domain*, interpretability is domain-specific and therefore eludes definitions that would be too restrictive. Understanding the decision process implies understanding the role of each parameter and feature and how they interact with each other to generate a prediction. We call this aspect the *transparency* of a predictor. Transparency is more easily obtained when the predictor is both *sparse* and *parsimonious*; that is, when the predictor has only a few non-zero parameters and when only a few features are retained by the predictor, respectively.

But interpretability is even more complicated than that. As pointed out by [86], characteristics such as monotonicity for a given variable, decomposability into sub-predictors, ability to perform case-based reasoning, usage of complex but well-known (e.g., Newton’s laws of physics) relations, preferences among the choice of variables, etc. impact just how interpretable a predictor is. Note that this notion is human-based and therefore is subjective, making it even harder to objectively quantify or define interpretability.

2.2 Explainability

All in all, interpretability greatly differs from *explaining* a predictor, where inherently hidden information concerning its decision process is presented to a given person. Explainability thus refers to the methods for creating these explanations. As stated by [8]: “[E]xplainability is associated with the notion of explanation as an interface between humans and a [predictor] that is, at the same time, both an accurate proxy of the [predictor] and comprehensible to humans.”

Those explanations are simplifications of the reality and, in machine learning, often presented in the form of simplifications of the original predictor [16,83,97,75,54,123,62]. Explainability techniques can be separated into several different (sometimes overlapping) subgroups, involving a varied nomenclature: *global* [7,38,118,41,25], describing the average behaviour of a predictor; *local* [35,59,104,15], describing a predictor behaviour for a given example; *model-agnostic* [71,26,14,94], applied to any model; *model-specific* [24,43,48,13,67], concerning a specific model; etc. This *explanatory* use of explainability is different

from its *exploratory* purpose [9] (i.e. being used practically during the training of ML predictors); for example, feature selection [111]. When conveying an explanation, the *explainer* is the person transmitting it and the *explainee* is the person receiving it.

When is the problem faced?	Number	The problem itself
Explainability	Problem 1	<i>There is an obligation to blindly trust the explanation.</i>
	Problem 2	<i>Explanations are necessarily wrong.</i>
	Problem 3	<i>There necessarily is a misalignment between what the explainer wants to convey and what the explainee actually understands.</i>
	Problem 4	<i>Explanations can create more questions than it answers.</i>
	Problem 5*	<i>Computing the explanations might involve huge computation time.</i>
	Problem 6*	<i>Most techniques are vulnerable to adversarial attacks.</i>
Interpretability	Problem 1	<i>Interpretability can't provide all that there is to know about a predictor.</i>
	Problem 2	<i>Interpretability is subjective.</i>
	Problem 3	<i>Enforcing interpretability can make training harder.</i>
	Problem 4	<i>Interpretability as a means is not enough to ensure some of its most common ends.</i>

Table 1: The principal limitations of explainability and interpretability that are discussed (raised*) in the article.

3 The relationship between interpretability and explainability

In this section, we make the case that interpretability and explainability actually are *complementary* to one another. In order to do so, we will look at the flaws and limitations in the information that can be gathered both in explaining black-boxes (Subsection 3.1) and in interpretable predictors that are not explained in any way (Subsection 3.2). The various points that are discussed are listed in Table 1. Then, we discuss how these problems can be attenuated when it comes to explained interpretable predictors (Subsection 3.3). Note that even though the flaws we will be discussing are raised separately, they must be understood as a whole, for they are not independent.

3.1 The Flaws in Explaining Black-Boxes

Problem E. 1 *There is an obligation to blindly trust the explanation.*

Explanations form a proxy between the user and something that is, by definition, inscrutable to human beings; they cannot be authentically verified but by other proxies. The explanations must therefore be trusted blindly, or substantiated with other explanations, where the same problem arises.

Some explainability techniques display interesting properties ensuring they behave as expected, but most of the time, and because explainability tends to be used on truly complex predictors, simplifications of the techniques (heuristics) are required, leading to cases where the properties do not even hold. This is notably the case for SHAP values [62] computed with the TreeSHAP algorithm [61], which violates the *missingness* property of SHAP values: "TreeSHAP was introduced as a fast, model-specific alternative to KernelSHAP, but it turned out that [...] features that have no influence on the prediction function f can get a TreeSHAP estimate different from zero" [72].

In addition, there is an obligation not only to trust the explanation but also the explainer. The explanations might truly reflect what the explainer wanted to convey, but this has to be aligned with what information the explainee wants to get. As stated in [10]: "most situations where explanations are requested are adversarial, meaning that the explanation provider and receiver have opposing interests and incentives, so that the provider might manipulate the explanation for her own ends."

Problem E. 2 *There necessarily is a misalignment between what the explainer wants to convey and what the explainee actually understands.*

Miller [70] gives precious insights on that matter: "[I]t is fair to say that most work in explainable artificial intelligence uses only the researchers' intuition of what constitutes a 'good' explanation. There exist vast and valuable bodies of research in philosophy, psychology, and cognitive science on how people define, generate, select, evaluate, and present explanations, which argues that people employ certain cognitive biases and social expectations towards the explanation process¹." Also: "[The explanations] are a transfer of knowledge, presented as part of a conversation or interaction, and are thus presented relative to the explainer's beliefs about the explainee's beliefs. This goes without discussing the misalignment between what the technique conveys and what the explainer actually wants to convey." This is notably the case for the TreeSHAP algorithm [61], which defines its value function differently than in the original work on SHAP values [62], thus changing the interpretation of the obtained values, but which could easily be misunderstood or ignored by the explainer.

Problems also arise on the explainee's part only: the various explainability techniques aiming at computing similar information (e.g. Permutation feature

¹ See Malle [63] when it comes to explanation selection, and Kahneman et al. [45], Miller and Gunasegaram [69] and Giroto et al. [34] when it comes to counterfactual explanations.

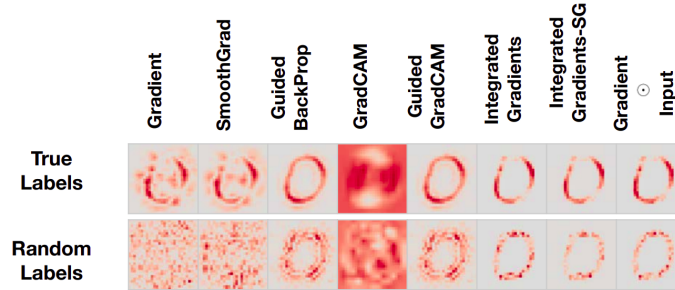


Fig. 1: Borrowed from [1], adapted: two predictors were trained on the MNIST task: one with the regular labels and one with random labels. Both predictors provide similar saliency maps for similar inputs even though we know for sure the second one hasn’t learned anything.

importance [25] and Leave one covariate out [58], both aiming at computing the same information) have been observed to yield important discrepancies, even on small tasks tackled by a simple sparse linear predictor [65]. If the subtleties distinguishing these methods are out of reach for most explainees, then how do explainees account for the discrepancies they observe?

Problem E. 3 *Explanations are necessarily wrong.*

The fact that explanations “condense the complexity of machine learning [predictors] into human-intelligible descriptions that only provide insight into specific aspects of the [predictor] and data” [73], or that “if the explanation was completely faithful to what the original [predictor] computes, the explanation would equal the original [predictor], and one would not need the original [predictor] in the first place, only the explanation” [85].

A black-box is likely to be a truly complex function (especially in the era of deep learning [99]) and while many explanation techniques rely on extrapolating between training examples [90,62,12,25,27,28,38], this can easily lead to unreliable results [74,42,73]. And even when explainability is used directly with training points, in many cases, assumptions are made on the data distribution (for example: assuming feature independence [62,83,96,92,18]) or the predictor itself (for example: assuming linearity of the predictor [62]). Such assumptions being most of the time unlikely to be met, this leads to unexpected results [72] or in worse cases misleading or false characterizations [55,85,53].

Human intuition can be fooled and therefore is unreliable for deciding to trust an explanation, which is sometimes right but for wrong, surprising, or incoherent reasons [1]. Figure 1 shows a speaking example of this behavior.

Problem E. 4 *Explanations can create more questions than it answers.*

Rudin et al. [86], among others [44], give an insightful example on that matter: “Saliency maps highlight the pixels of an image that are used for a prediction,

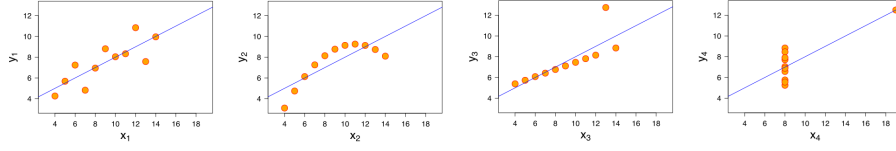


Fig. 2: Anscombe's quartet [6].

but they do not explain how the pixels are used. As an analogy, consider a real estate agent who is pricing a house. A *black-box* real estate agent would provide the price with no explanation. A *saliency* real estate agent would say that the price is determined from the roof and backyard, but doesn't explain how the roof and backyard were used to determine the price."

Figure 1 shows an example where anyone could get fooled by assuming both predictors to have learned how to recognize the shape of a "0" digit, but only knowledge of the training scheme reveals that it is not the case. Here are two other problems (among others) that won't be discussed in detail, but that are worth raising:

- *Computing the explanations might involve huge computation time*; As raised in Problem E. 1, being faithful to the proposed in the literature might be a computational burden [25,107,62], leading to approximations and heuristics which in turn can lead to Problem E. 2;
- *Explainability techniques are vulnerable to adversarial attacks*; not only when it comes to the values that are computed [53,55,93,31,52], but also in the sense instilled in the obtained values [10,103].

3.2 The Flaws of Interpretable Predictors

Problem I. 1 *Interpretability can't provide all that there is to know about a predictor.*

Anscombe's quartet [6] (see Figure 2) efficiently illustrates that many datasets, given a training algorithm, could lead to a same predictor: "[they] are designed to have approximately the same linear regression line (as well as nearly identical means, standard deviations, and correlations) but are graphically very different."² This illustrates the pitfalls of relying solely on a fitted predictor for trying to understand the relationship between variables or the role of each of its parameters.

Many other information aren't displayed by the predictor in itself: adversarial examples, or how is it possible to trick the decision process; influential instances, or how influential was a certain training example when training the predictor; feature interaction, or to what extent the prediction is the result of joint effects of the features; relative feature importance; counterfactual explanations, or how an instance has to change to significantly change its prediction; etc.

² https://en.wikipedia.org/wiki/Linear_regression

Problem I. 2 *Interpretability is subjective.*

Interpretability cannot be objectively computed, for it depends on both how the judge makes sense of the predictor and their preferences: it is an opinion. Therefore, two judges could have a completely different idea of the inner workings of a given predictor. And as for opinion, one can only argue in favor of his views on the interpretability of a predictor, but can't acquire a view on a predictor from someone else.

Problem I. 3 *Enforcing interpretability can make training harder.*

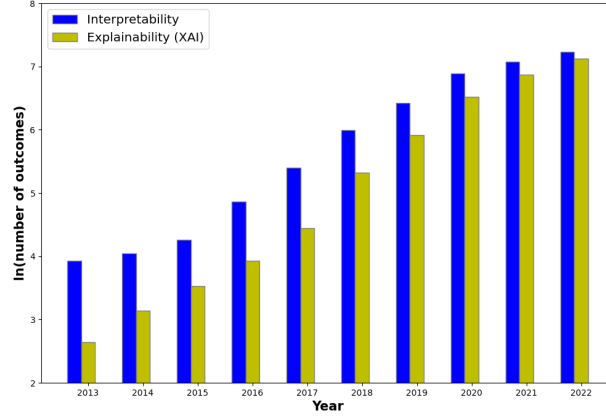
Since the space of interpretable predictors is a subspace of the ensemble of every predictor, enforcing interpretability is a constraint to apply to an algorithm. In practice, the following are usually considered for favoring transparency: reducing the number of features of the predictor, the number of non-zero parameters of the predictor, or the number of bits required in order to encode some of the predictor's parameters. Such constraints can lead to delicate discrete optimization problems where making steps forward in efficiently handling those is not an easy task. For instance, when it comes to the efficient training of decision trees, CART [11] (which dates back to 1984) still is a state-of-the-art approach for training decision trees.

This might explain the trend presented on Figure 3, where it seems like there is a tendency, probably in agreement with the idea that explainability and interpretability are substitutes for one another, to neglect interpretable approaches and favoring the training of black-boxes before explaining them.

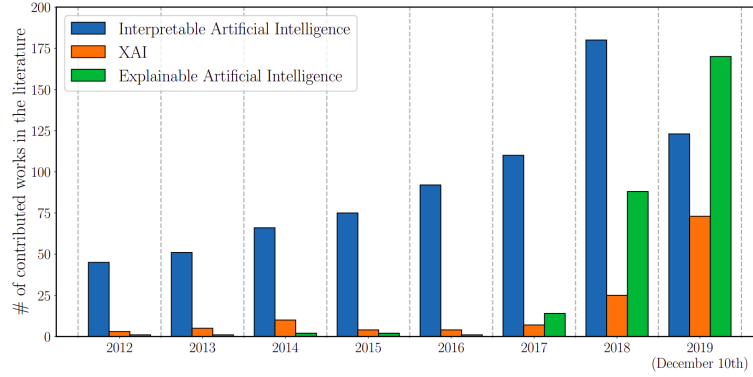
Problem I. 4 *Interpretability as a means is not enough to ensure some of its most common ends.*

It is clear that interpretability is needed for many reasons: if algorithms lack transparency, domain experts or the public will not trust them [46]; people have a right to know why an ML algorithm has produced some verdict (such as lack of creditworthiness) about them [36]; algorithms whose inner workings are not interpretable can't enable us to produce causal explanations of the world [82]; etc.

For example, fairness is an end to which interpretability is sometimes used as a mean [19,72], where the predictions need to be unbiased toward one or some sensitive variables. Not using such variables in the decision process simply isn't enough to ensure fairness because sometimes, information regarding the sensible variables is encapsulated in benign variables via complex relationships. Such relationships are unknown and might be numerous. For these reasons, the different aspects of fairness (e.g. independence, separation, sufficiency) are mostly only statistically demonstrated. Interpretability only permits the knowing of exactly how the benign variables are manipulated by a predictor, which simply is not enough to certify its fairness [3]. This goes as well for privacy, where sensible variables would be personal information.



(a) Comparison between the natural log of the number of outcomes for specific query searches on ArXiv as a function of the year. Searches were done for the Computer Science (cs) subject, with cross-listed papers. The query for Interpretability: ("Interpretability" $\vee$ "Interpretable" in Title) $\vee$ ("Interpretability machine learning" $\vee$ "Interpretable machine learning" in Abstract). Query for Explainability: ("Explainability" $\vee$ "Explainable" $\vee$ "XAI" in Title) $\vee$ ("Explainability machine learning" $\vee$ "Explainable machine learning" $\vee$ "XAI" in Abstract). One must not forget that in many cases, "interpretability" refers to our definition of "explainability", but the inverse is not true.



(b) Borrowed from [8]: "Evolution of the number of total publications whose title, abstract and/or keywords refer to the field of XAI during the last years. Data retrieved from Scopus® (December 10th, 2019) by using the search terms indicated in the legend when querying this database. [...]"

Fig. 3: Trends in interpretability and explainability (XAI) research popularity.

Reliability, robustness or trust are also ends to which interpretability is sometimes used as a mean, each of them being seemingly connected: can we trust a predictor that is not reliable? Can it be reliable if it is not robust? The problem is that interpretability does not imply any of these characteristics. A predictor that is interpretable is not necessarily a predictor in which one can blindly put its trust [91]: interpretability only yields an argument, which can be used to back the proposition that we should trust (or not) the predictor: it yields clarity, not a guarantee. That same predictor would not necessarily be robust or reliable either, but understanding the predictor and its training algorithm leads to insights into its limitations and weaknesses. In the end, interpretability induces confidence in how much a predictor *can* be reliable, robust, or trusted.

3.3 Explained Interpretable Predictor

Here, we make the case that interpretability and explainability are not substitutes for one another, but in many occasions complementary, in the sense that they reinforce one another. We do so by looking at all of the problems that have been raised in Subsection 3.1 and Subsection 3.2, but now when explainability and interpretability meet.

First, the blindness with which the explainability approaches must be trusted (Problem E. 1) substantially decreases, for some explanations can be more easily verified or corroborated by a look at the interpretable predictor itself. Interpretability permits us to justify whether or not to put our trust in an explanation. The same goes for the over-reliance on explainability. The over-reliance might exist partially because the explanations are the only resources we have for understanding what is going on within the predictor. With interpretable predictors, it is now easier either to be skeptical toward explanations or to justify this reliance.

The fact that each explainability technique is, to a certain extent, misaligned either with the explainer’s intent or the explainee’s (Problem E. 2) is also attenuated. With a black-box, the explanation has to be received and analyzed by the explainee with no reference whatsoever, or other explanations. With an interpretable predictor, there is an objective common ground on which both the explainer and the explainee can understand each other; indeed, having a reference point when it comes to transferring knowledge (explanation) makes the chances of a discrepancy between what was intended to share and what has been perceived less likely to happen, and of a smaller scale when it does happen [70]. As for the discrepancies between explainability techniques, [65] shows that for a decreasing number of features, there is a decreasing disagreement between different feature importance techniques. The number of considered features (parsimony) is only a facet of interpretability, but this might be a first step toward demonstrating that the more interpretable the predictor is, the fewer discrepancies there are between the explainer’s intent and the explainee’s interpretation of the information.

Even though explanations are always wrong (Problem E. 3), they are less wrong when it comes to interpretable predictors. Indeed, since explanations are simplifications of the original predictor, the simpler the original predictor (which

in many cases is proportionate to the degree of interpretability), the lesser of an approximation is required by the explanation. Thus, the explanations might just be more faithful to the original predictor. In the same line of thought, with an interpretable predictor, there are many cases where the computational costs will be reduced, thus permitting the use of more computationally heavy but also more authentic methods [62].

In some cases, the explanations do raise questions themselves (Problem E. 4), but the interpretable predictor can answer those ones. For example, an attention-based technique [106,89,115] aims at identifying what input has been mostly important for rendering a given prediction, but raises the following question: “How did those inputs were actually used to render that prediction?” With an interpretable predictor, that later question inherently can be answered. We mentioned earlier in Problem E. 4 the example from [85] of real estate agents who are pricing a house, where “a *black-box* real estate agent would provide the price with no explanation”, whereas “a saliency real estate agent would say that the price is determined from the roof and backyard, but does not explain how they were used to determine the price”. Well, in contrast, “an interpretable agent would explain the calculation in detail, for instance, using *comps* or comparable properties to explain how the roof and backyard are comparable between properties, and how these comparisons were used to determine the price”.

Finally, simpler (parsimonious, sparse) predictors tend to be more robust toward adversarial attacks [100,77,78,84]. For example, [78] propose a framework for distilling networks in order to make them more robust; one could think of a scenario where the distillation is made after a disentanglement layer in a deep neural network. [84] shows similar conclusions, but this time in a practical context: when the inputs are medical images.

We made the case that interpretability can help reduce the flaws in explaining a predictor. Here, we make the case that the flaws of interpretability, when applied to explained predictors, are attenuated.

Interpretability is not sufficient to provide information about the parameters and features of a predictor (Problem I. 1), whereas various explainability techniques can compute feature importance, distribution patterns, visual representation of the inputs/features and the predictor, statistics on parameters and inputs, etc. By doing so, humans don’t have to let their intuition guide them in these matters, so it leads both the judges toward a single view on the predictor (objectivity replaces subjectivity, thus diminishing Problem I. 2).

Even though the training of interpretable predictors can be harder than the training of any kind of predictor (Problem I. 3), explainability, in its exploratory purpose, can come in handy: variable selection [111,49,117], parameter pruning [109,98], guiding the modeling of the predictor [108] or understanding the causal relationship between variables [112] all can help during the training phase. For example, in order to favor parsimony, one could build its predictor in a top-down fashion but make use of SHAP values for deciding which features to remove.

Interpretability alone isn't enough to ensure some of its most common ends (Problem I. 4). When it comes to group fairness (independence, separation, sufficiency) or individual fairness, interpretability might ensure that the benign variables are used properly, but explainability only is able to compute metrics assessing true fairness [120,95,121] and provide many relevant statistics (feature importance, distribution patterns, etc.) for guaranteeing the reliability or trustworthiness of a predictor. Explainability can even be used in an adversarial fashion to find examples assessing the biases of a predictor.

4 Conclusion

We studied the major limitations of a predictor involving only interpretability or explainability and we conclude that considering both these aspects simultaneously leads to the mitigation of these many flaws: being less subjective in the interpretation of the predictor; ensuring that the predictor meets his ends; more reliable and truthful information; etc. Our analysis leads to a better understanding of these two abstract concepts in themselves and in their relationship. We invite the ML community to look at interpretability and explainability in a new way, and we hope it fosters ideas where both interpretability and explainability are met at the same time.

Similar to what motivated the present work, it would be important to put forth the inquiry on interpretability and explainability and empirically verify the benefits of being both interpretable and explainable on the following topics: robustness [40,88,105], algorithmic stability [80,101,47], fairness [2,22,114] / unbiasedness [29,5,119], private life/data privacy [110,113,122], etc.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I.J., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: NeurIPS. pp. 9525–9536 (2018)
2. Agarwal, A., Agarwal, H., Agarwal, N.: Fairness score and process standardization: framework for fairness certification in artificial intelligence systems. *AI Ethics* **3**(1), 267–279 (2023)
3. Agarwal, S.: Trade-offs between fairness and interpretability in machine learning. In: IJCAI 2021 Workshop on AI for Social Good (2021)
4. Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K.V., Hoos, H.H., Hung, H., Jonker, C.M., Monz, C., Neerincx, M.A., Oliehoek, F.A., Prakken, H., Schlobach, S., van der Gaag, L.C., van Harmelen, F., van Hoof, H., van Riemsdijk, B., van Wynsberghe, A., Verbrugge, R., Verheij, B., Vossen, P., Welling, M.: A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer* **53**(8), 18–28 (2020)
5. Amrani, N., Serra-Sagristà, J., Marcellin, M.W.: Unbiasedness of regression wavelet analysis for progressive lossy-to-lossless coding. In: PCS. pp. 1–5. IEEE (2016)
6. Anscombe, F.J.: Graphs in statistical analysis. *The American Statistician* **27**(1), 17–21 (1973), <http://www.jstor.org/stable/2682899>

7. Apley, D.W., Zhu, J.: Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82** (2016)
8. Arrieta, A.B., Rodríguez, N.D., Ser, J.D., Benetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020)
9. Atrey, A., Clary, K., Jensen, D.D.: Exploratory not explanatory: Counterfactual analysis of saliency maps for deep reinforcement learning. In: ICLR. OpenReview.net (2020)
10. Bordt, S., Finck, M., Raidl, E., von Luxburg, U.: Post-hoc explanations fail to achieve their purpose in adversarial contexts. In: FAccT. pp. 891–905. ACM (2022)
11. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: *Classification and Regression Trees*. Wadsworth (1984)
12. Casalicchio, G., Molnar, C., Bischl, B.: Visualizing the feature importance for black box models. In: ECML/PKDD (1). *Lecture Notes in Computer Science*, vol. 11051, pp. 655–670. Springer (2018)
13. da Costa F. Chaves, A., Vellasco, M.M.B.R., Tanscheit, R.: Fuzzy rule extraction from support vector machines. In: HIS. pp. 335–340. IEEE Computer Society (2005)
14. Dabkowski, P., Gal, Y.: Real time image saliency for black box classifiers. In: NIPS. pp. 6967–6976 (2017)
15. Dandl, S., Molnar, C., Binder, M., Bischl, B.: Multi-objective counterfactual explanations. In: PPSN (1). *Lecture Notes in Computer Science*, vol. 12269, pp. 448–469. Springer (2020)
16. Das, A., Rad, P.: Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *CoRR* **abs/2006.11371** (2020)
17. Dastin, J.: Amazon scraps secret ai recruiting tool that showed bias against women. Reuters (10 2018), <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>
18. Datta, A., Sen, S., Zick, Y.: Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In: IEEE Symposium on Security and Privacy. pp. 598–617. IEEE Computer Society (2016)
19. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017), <http://arxiv.org/abs/1702.08608>, cite arxiv:1702.08608
20. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning* (2017)
21. Drissi, M., Watkins, O., Khant, A., Ojha, V., Sandoval, P., Segev, R., Weiner, E., Keller, R.: Program Language Translation Using a Grammar-Driven Tree-to-Tree Model. *arXiv e-prints* arXiv:1807.01784 (Jul 2018). <https://doi.org/10.48550/arXiv.1807.01784>
22. Fabris, A., Esuli, A., Moreo, A., Sebastiani, F.: Measuring fairness under unawareness of sensitive attributes: A quantification-based approach. *J. Artif. Intell. Res.* **76**, 1117–1180 (2023)
23. Fan, F.L., Xiong, J., Li, M., Wang, G.: On interpretability of artificial neural networks: A survey. *IEEE Transactions on Radiation and Plasma Medical Sciences* **5**(6), 741–760 (2021). <https://doi.org/10.1109/TRPMS.2021.3066428>
24. Féraud, R., Clérot, F.: A methodology to explain neural network classification. *Neural Networks* **15**(1), 237–246 (2002)

25. Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* **20**, 177:1–177:81 (2019)
26. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: *ICCV*. pp. 3449–3457. IEEE Computer Society (2017)
27. Friedman, J.H.: Multivariate Adaptive Regression Splines. *The Annals of Statistics* **19**(1), 1 – 67 (1991). <https://doi.org/10.1214/aos/1176347963>, <https://doi.org/10.1214/aos/1176347963>
28. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29**(5), 1189 – 1232 (2001). <https://doi.org/10.1214/aos/1013203451>, <https://doi.org/10.1214/aos/1013203451>
29. Friedrich, T., Kötzing, T., Krejca, M.S.: Unbiasedness of estimation-of-distribution algorithms. *Theor. Comput. Sci.* **785**, 46–59 (2019)
30. Gacto, M.J., Alcalá, R., Herrera, F.: Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Inf. Sci.* **181**(20), 4340–4360 (2011)
31. Ghorbani, A., Abid, A., Zou, J.Y.: Interpretation of neural networks is fragile. In: *AAAI*. pp. 3681–3688. AAAI Press (2019)
32. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M.A., Kagal, L.: Explaining explanations: An approach to evaluating interpretability of machine learning. *CoRR* **abs/1806.00069** (2018)
33. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M.A., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: *DSAA*. pp. 80–89. IEEE (2018)
34. Girotto, V., Legrenzi, P., Rizzo, A.: Event controllability in counterfactual thinking. *Acta Psychologica* **78**(1), 111–133 (1991). [https://doi.org/https://doi.org/10.1016/0001-6918\(91\)90007-M](https://doi.org/https://doi.org/10.1016/0001-6918(91)90007-M), <https://www.sciencedirect.com/science/article/pii/000169189190007M>
35. Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics* **24**, 44 – 65 (2013)
36. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a “right to explanation”. *AI Magazine* **38**(3), 50–57 (Oct 2017). <https://doi.org/10.1609/aimag.v38i3.2741>, <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2741>
37. Graziani, M., Dutkiewicz, L., Calvaresi, D., Amorim, J.P., Yordanova, K., Vered, M., Nair, R., Abreu, P.H., Blanke, T., Pulignano, V., Prior, J.O., Lauwaert, L., Reijers, W., Depeursinge, A., Andrearczyk, V., Müller, H.: A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. *Artif. Intell. Rev.* **56**(4), 3473–3504 (2023)
38. Greenwell, B.M., Boehmke, B.C., McCarthy, A.J.: A simple and effective model-based variable importance measure. *CoRR* **abs/1805.04755** (2018)
39. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. *ACM Comput. Surv.* **51**(5), 93:1–93:42 (2019)
40. Hia, S., Kuswanto, H., Prastyo, D.D.: Robustness of support vector regression and random forest models: A simulation study. In: *DaSET. Lecture Notes on Data Engineering and Communications Technologies*, vol. 165, pp. 465–479. Springer (2022)
41. Hooker, G.: Discovering additive structure in black box functions. In: *KDD*. pp. 575–580. ACM (2004)

42. Hooker, G., Mentch, L., Zhou, S.: Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Stat. Comput.* **31**(6), 82 (2021)
43. Hu, J., Cao, L., Tong, T., Ye, Q., Zhang, S., Li, K., Huang, F., Shao, L., Ji, R.: Architecture disentanglement for deep neural networks. In: *ICCV*. pp. 652–661. *IEEE* (2021)
44. Jain, S., Wallace, B.C.: Attention is not explanation. In: *NAACL-HLT* (1). pp. 3543–3556. Association for Computational Linguistics (2019)
45. Kahneman, D., Slovic, P., Tversky, A. (eds.): *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press (1982)
46. Kim, B.: Interactive and interpretable machine learning models for human machine collaboration (2015), <https://api.semanticscholar.org/CorpusID:110496388>
47. Kim, B., Barber, R.F.: Black box tests for algorithmic stability. *CoRR abs/2111.15546* (2021)
48. Kindermans, P., Schütt, K.T., Alber, M., Müller, K., Erhan, D., Kim, B., Dähne, S.: Learning how to explain neural networks: Patternnet and patternattribution. In: *ICLR* (Poster). *OpenReview.net* (2018)
49. Krause, J., Perer, A., Bertini, E.: INFUSE: interactive feature selection for predictive modeling of high dimensional data. *IEEE Trans. Vis. Comput. Graph.* **20**(12), 1614–1623 (2014)
50. Krishnan, M.: Against interpretability: A critical examination of the interpretability problem in machine learning. *Philosophy and Technology* **33**(3), 487–502 (2020). <https://doi.org/10.1007/s13347-019-00372-9>
51. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) *Advances in Neural Information Processing Systems*. vol. 25. Curran Associates, Inc. (2012), https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
52. Laberge, G., Aïvodji, U., Hara, S.: Fooling SHAP with stealthily biased sampling. *CoRR abs/2205.15419* (2022)
53. Lakkaraju, H., Bastani, O.: "how do I fool you?": Manipulating user trust via misleading black box explanations. In: *AIES*. pp. 79–85. *ACM* (2020)
54. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Interpretable & explorable approximations of black box models. *CoRR abs/1707.01154* (2017)
55. Laugel, T., Lesot, M., Marsala, C., Renard, X., Detyniecki, M.: The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In: *IJCAI*. pp. 2801–2807. *ijcai.org* (2019)
56. Leavitt, M.L., Morcos, A.S.: Towards falsifiable interpretability research. *CoRR abs/2010.12016* (2020)
57. Leblanc, B., Germain, P.: A greedy algorithm for building compact binary activated neural networks. *CoRR abs/2209.03450* (2022)
58. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L.: Distribution-free predictive inference for regression. *Journal of the American Statistical Association* **113**(523), 1094–1111 (2018). <https://doi.org/10.1080/01621459.2017.1307116>, <https://doi.org/10.1080/01621459.2017.1307116>
59. Linsley, D., Scheibler, D., Eberhardt, S., Serre, T.: Global-and-local attention networks for visual recognition. *CoRR abs/1805.08819* (2018)
60. Lipton, Z.C.: The mythos of model interpretability. *Commun. ACM* **61**(10), 36–43 (2018)
61. Lundberg, S.M., Erion, G.G., Lee, S.: Consistent individualized feature attribution for tree ensembles. *CoRR abs/1802.03888* (2018)

62. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: NIPS. pp. 4765–4774 (2017)
63. Malle, B.: How the mind explains behavior: Folk explanation, meaning and social interaction (01 2004)
64. Marchand, M., Shawe-Taylor, J.: Learning with the set covering machine. In: ICML. pp. 345–352. Morgan Kaufmann (2001)
65. Markus, A.F., Fridgeirsson, E.A., Kors, J.A., Verhamme, K.M., Reps, J.M., Rijnbeek, P.R.: Understanding the size of the feature importance disagreement problem in real-world data. In: ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH) (2023), <https://openreview.net/forum?id=FKjFUEV63f>
66. Markus, A.F., Kors, J.A., Rijnbeek, P.R.: The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J. Biomed. Informatics* **113**, 103655 (2021)
67. Martens, D., Baesens, B., Gestel, T.V., Vanthienen, J.: Comprehensible credit scoring models using rule extraction from support vector machines. *Eur. J. Oper. Res.* **183**(3), 1466–1476 (2007)
68. Martinez, E., Kirchner, L.: The secret bias hidden in mortgage-approval algorithms. *The Markup* (08 2021), <https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms>
69. Miller, D.T., Gunasegaram, S.: Temporal order and the perceived mutability of events: Implications for blame assignment. *Journal of Personality and Social Psychology* (Dec 1990). <https://doi.org/10.1037/0022-3514.59.6.1111>
70. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **267**, 1–38 (2019)
71. Mishra, S., Sturm, B.L., Dixon, S.: Local interpretable model-agnostic explanations for music content analysis. In: ISMIR. pp. 537–543 (2017)
72. Molnar, C.: Interpretable machine learning (2022), <https://christophm.github.io/interpretable-ml-book>
73. Molnar, C., König, G., Herbringer, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., Casalicchio, G., Grosse-Wentrup, M., Bischl, B.: General pitfalls of model-agnostic interpretation methods for machine learning models. In: *xxAI@ICML. Lecture Notes in Computer Science*, vol. 13200, pp. 39–68. Springer (2020)
74. Molnar, C., König, G., Bischl, B., Casalicchio, G.: Model-agnostic feature importance and effects with dependent features - A conditional subgroup approach. *CoRR* **abs/2006.04628** (2020)
75. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition* **65**, 211–222 (2017). <https://doi.org/https://doi.org/10.1016/j.patcog.2016.11.008>, <https://www.sciencedirect.com/science/article/pii/S0031320316303582>
76. Naji, M.A., Filali, S.E., Aarika, K., Benlahmar, E.H., Abdelouahid, R.A., Debauche, O.: Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia Computer Science* **191**, 487–492 (2021). <https://doi.org/https://doi.org/10.1016/j.procs.2021.07.062>, <https://www.sciencedirect.com/science/article/pii/S1877050921014629>, the 18th International Conference on Mobile Systems and Pervasive Computing (MobiSPC), The 16th International Conference on Future Networks and Communications (FNC), The 11th International Conference on Sustainable Energy Information Technology

77. Papernot, N., McDaniel, P.D., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: EuroS&P. pp. 372–387. IEEE (2016)
78. Papernot, N., McDaniel, P.D., Wu, X., Jha, S., Swami, A.: Distillation as a defense to adversarial perturbations against deep neural networks. In: IEEE Symposium on Security and Privacy. pp. 582–597. IEEE Computer Society (2016)
79. Paulus, R., Socher, R., Manning, C.D.: Global belief recursive neural networks. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) *Advances in Neural Information Processing Systems*. vol. 27. Curran Associates, Inc. (2014), https://proceedings.neurips.cc/paper_files/paper/2014/file/1415db70fe9ddb119e23e9b2808cde38-Paper.pdf
80. Raj, A., Barsbey, M., Gürbüzbalaban, M., Zhu, L., Simsekli, U.: Algorithmic stability of heavy-tailed stochastic gradient descent on least squares. In: ALT. *Proceedings of Machine Learning Research*, vol. 201, pp. 1292–1342. PMLR (2023)
81. Ras, G., Xie, N., van Gerven, M., Doran, D.: Explainable deep learning: A field guide for the uninitiated. *J. Artif. Intell. Res.* **73**, 329–396 (2022). <https://doi.org/10.1613/JAIR.1.13200>, <https://doi.org/10.1613/jair.1.13200>
82. Ratti, E., López-Rubio, E.: Mechanistic models and the explanatory limits of machine learning (2018), <http://philsci-archive.pitt.edu/14452/>
83. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: KDD. pp. 1135–1144. ACM (2016)
84. Rodriguez, D., Nayak, T., Chen, Y., Krishnan, R., Huang, Y.: On the role of deep learning model complexity in adversarial robustness for medical images. *BMC Medical Informatics Decis. Mak.* **22-S(2)**, 160 (2022)
85. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1(5)**, 206–215 (2019)
86. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C.: Interpretable machine learning: Fundamental principles and 10 grand challenges (2021)
87. Saeed, W., Omlin, C.W.: Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowl. Based Syst.* **263**, 110273 (2023)
88. Saengsawang, S., Li, G.: Theoretical analysis of norm selection for robustness verification of neural networks. *Phys. Commun.* **58**, 102019 (2023)
89. Seo, S., Huang, J., Yang, H., Liu, Y.: Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In: RecSys. pp. 297–305. ACM (2017)
90. Shapley, L.S.: 17. A Value for n-Person Games, pp. 307–318. Princeton University Press, Princeton (1953). <https://doi.org/doi:10.1515/9781400881970-018>, <https://doi.org/10.1515/9781400881970-018>
91. Shen, M.W.: Trust in AI: interpretability is not necessary or sufficient, while black-box interaction is necessary and sufficient. *CoRR* **abs/2202.05302** (2022)
92. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: ICML. *Proceedings of Machine Learning Research*, vol. 70, pp. 3145–3153. PMLR (2017)
93. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: AIES. pp. 180–186. ACM (2020)
94. Staniak, M., Biecek, P.: Explanations of model predictions with live and breakdown packages. *R J.* **10(2)**, 395 (2018)

95. Stevens, A., Deruyck, P., Veldhoven, Z.V., Vanthienen, J.: Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva. In: SSCI. pp. 1241–1248. IEEE (2020)
96. Strumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* **41**(3), 647–665 (2014)
97. Subramanian, A., Pruthi, D., Jhamtani, H., Berg-Kirkpatrick, T., Hovy, E.H.: SPINE: sparse interpretable neural embeddings. In: AAAI. pp. 4921–4928. AAAI Press (2018)
98. Sudars, K., Namatēvs, I., Ozols, K.: Improving performance of the prystine traffic sign classification by using a perturbation-based explainability approach. *Journal of Imaging* **8**(2) (2022). <https://doi.org/10.3390/jimaging8020030>, <https://www.mdpi.com/2313-433X/8/2/30>
99. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR. pp. 1–9. IEEE Computer Society (2015)
100. Tabacof, P., Valle, E.: Exploring the space of adversarial images. In: IJCNN. pp. 426–433. IEEE (2016)
101. Tamar, A., Soudry, D., Zisselman, E.: Regularization guarantees generalization in bayesian reinforcement learning through algorithmic stability. In: AAAI. pp. 8423–8431. AAAI Press (2022)
102. Telford, T.: Apple card algorithm sparks gender bias allegations against goldman sachs. *The Washington Post* (11 2019), <https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/>
103. Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., Paoli, J., Puig, S., Rosendahl, C., Soyer, H.P., Zalaudek, I., Kittler, H.: Human–computer collaboration for skin cancer recognition. *Nat Med* **26**, 1229–1234 (06/2020 2020). <https://doi.org/10.1038/s41591-020-0942-0>
104. Wachter, S., Mittelstadt, B.D., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR abs/1711.00399* (2017)
105. Wang, B., Pang, M., Dong, Y.: Turning strengths into weaknesses: A certified robustness inspired attack framework against graph neural networks. *CoRR abs/2303.06199* (2023)
106. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: CVPR. pp. 6450–6458. IEEE Computer Society (2017)
107. Wei, P., Lu, Z., Song, J.: Variable importance analysis: A comprehensive review. *Reliab. Eng. Syst. Saf.* **142**, 399–432 (2015). <https://doi.org/10.1016/J.RESS.2015.05.018>, <https://doi.org/10.1016/j.ress.2015.05.018>
108. Yang, Z., Zhang, A., Sudjianto, A.: Enhancing explainability of neural networks through architecture constraints. *IEEE Trans. Neural Networks Learn. Syst.* **32**(6), 2610–2621 (2021)
109. Yeom, S.K., Seegerer, P., Lapuschkin, S., Binder, A., Wiedemann, S., Müller, K.R., Samek, W.: Pruning by explaining: A novel criterion for deep neural network pruning. *Pattern Recognition* **115**, 107899 (2021). <https://doi.org/https://doi.org/10.1016/j.patcog.2021.107899>, <https://www.sciencedirect.com/science/article/pii/S0031320321000868>

110. Yu, S., Cui, L.: Security and Privacy in Federated Learning. Springer (2023)
111. Zacharias, J., von Zahn, M., Chen, J., Hinz, O.: Designing a feature selection method based on explainable artificial intelligence. *Electron. Mark.* **32**(4), 2159–2184 (2022)
112. Zednik, C., Boelsen, H.: The exploratory role of explainable artificial intelligence (2020), <http://philsci-archive.pitt.edu/18005/>
113. Zhang, J., Zhou, J., Guo, J., Sun, X.: Visual object detection for privacy-preserving federated learning. *IEEE Access* **11**, 33324–33335 (2023)
114. Zhang, T., Zhu, T., Han, M., Chen, F., Li, J., Zhou, W., Yu, P.S.: Fairness in graph-based semi-supervised learning. *Knowl. Inf. Syst.* **65**(2), 543–570 (2023)
115. Zhang, X., Wang, Z.: Spatial proximity feature selection with residual spatial-spectral attention network for hyperspectral image classification. *IEEE Access* **11**, 23268–23281 (2023)
116. Zhang, Y., Tiño, P., Leonardis, A., Tang, K.: A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence* **5**, 726–742 (2020)
117. Zhao, J., Karimzadeh, M., Masjedi, A., Wang, T., Zhang, X., Crawford, M.M., Ebert, D.S.: Featureexplorer: Interactive feature selection and exploration of regression models for hyperspectral images. In: *IEEE VIS (Short Papers)*. pp. 161–165. IEEE (2019)
118. Zhao, Q., Hastie, T.: Causal interpretations of black-box models. *Journal of Business & Economic Statistics* **39**(1), 272–281 (2021). <https://doi.org/10.1080/07350015.2019.1624293>, <https://doi.org/10.1080/07350015.2019.1624293>
119. Zhao, Y., He, X., Ma, L., Liu, H.: Unbiasedness-constrained least squares state estimation for time-varying systems with missing measurements under round-robin protocol. *Int. J. Syst. Sci.* **53**(9), 1925–1941 (2022)
120. Zhao, Y., Wang, Y., Derr, T.: Fairness and explainability: Bridging the gap towards fair model explanations. In: *AAAI*. pp. 11363–11371. AAAI Press (2023)
121. Zhou, J., Chen, F., Holzinger, A.: Towards explainability for AI fairness. In: *xxAI@ICML. Lecture Notes in Computer Science*, vol. 13200, pp. 375–386. Springer (2020)
122. Zhou, M., Zheng, Y., Wang, S., Hua, Z., Huang, H., Gao, Y., Jia, X.: PPTA: A location privacy-preserving and flexible task assignment service for spatial crowdsourcing. *Comput. Networks* **224**, 109600 (2023)
123. Zhou, Y., Hooker, G.: Interpreting models via single tree approximation. *arXiv: Methodology* (2016)
124. Zytek, A., Arnaldo, I., Liu, D., Berti-Équille, L., Veeramachaneni, K.: The need for interpretable features: Motivation and taxonomy. *SIGKDD Explor.* **24**(1), 1–13 (2022)